ANALYSIS OF SPARSE ELIMINATION

James R. Bunch

Cornell University

Prepared for:

Office of Naval Research

January 1973

ANALYSIS OF SPARSE ELIMINATION

James R. Bunch

Cornell University

TR 73-158

January, 1973

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Cornell University	unclassified
	2b. GROUP

3. REPORT TITLE

Analysis of Sparse Elimination

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report

5. AUTHOR(S) (First name, middle initial, last name)

James R. Bunch

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
January 1973	17	5

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-67-A-0077-0021	
b. PROJECT NO.	73-158
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT

Approved for public release, distribution unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Office of Naval Research, Rochester

13. ABSTRACT

An error analysis is presented for Gaussian elimination when the matrix is arbitrarily sparse. Error analyses for elimination on band matrices and full matrices follow as special cases.

DD FORM 1473 (PAGE 1)

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
sparse matrix						
error analysis						
Gaussian elimination						
band matrix						
graph theory						

1. INTRODUCTION.

Since the direct solution of systems of linear equations by elimination is now well understood when the matrix is full or band, attention has turned to the problem of elimination of systems when the matrix is arbitrarily sparse. When the sparse matrix arises from the discretization of an elliptic partial differential equation, it has a special structure and special properties which make iterative methods attractive. Even here, however, one can often find an ordering of the equations which makes elimination competitive (George[3]). But all too frequently the matrix has an arbitrarily sparse structure, e.g. in network analysis and econometric problems.

Here we present an error analysis of (point) Gaussian elimination when the matrix is arbitrarily sparse. The analysis is presented in terms of the structure of the elimination graphs arising in the elimination process. The only assumption we make is that the leading principal minors are non-zero, i.e. that the (1,1) entry in each reduced matrix is non-zero (cf. Forsythe and Moler [2], pp. 27-36).

Since the graphs of band and full matrices are special elimination graphs, the error analyses for band and full matrices follow as special cases.

2. MATRICES AND GRAPHS.

Let M be an n×n matrix with non-zero leading principal minors; hence M is non-singular and the LU decomposition of M exists and is unique. If M is symmetric, then $M=LDL^t$.

Rose [4] associated an undirected graph with a symmetric matrix and interpreted the LDL^t factorization graph-theoretically by undirected elimination graphs. Bunch and Rose [1] made the extension of the association of directed graphs with general square matrices and of the interpretation of the LU factorization by directed elimination graphs.

The directed graph of M, $G(M)=(X,\mathcal{A})$, with vertex set X and arc set $\mathcal{A}$, is defined as follows: a vertex $x_i \in X$ is associated with row i of M, and $(x_i,x_j) \in \mathcal{A}$ (an arc from x_i to x_j is in G) if and only if $m_{ij} \neq 0$ and $i \neq j$. The vertices X are regarded as ordered; i.e., $X=\{x_i\}_{i=1}^n$.

Let the matrix M be written as $M=\begin{bmatrix} a & r^t \\ c & B \end{bmatrix}$, where a is 1×1, r and c are (n-1)×1 and B is (n-1)×(n-1). Then the first step of the LU factorization of M can be written as

$$M=\begin{bmatrix} 1 & 0 \\ c/a & I \end{bmatrix} \qquad \begin{bmatrix} a & r^t \\ 0 & B-cr^t/a \end{bmatrix} .$$

If G(M) is the directed graph of M, the elimination graph G_y is obtained from G by deleting y together with its incident arcs and adding an arc (x,z) whenever there exists a (directed) x, z path of length 2 containing y. G_y is the graph of the matrix obtained by "eliminating" the variable corresponding to y in Gaussian elimination; e.g., G_{x_1} is the graph of $B-cr^t/a$. The accidental creation of zeros during the elimination process is ignored. For a more detailed discussion, see Bunch and Rose [1], Section 2.

Let $G_1,\ldots,G_{n-1}$ be the sequence of elimination graphs defined recursively by $G_0=G(M)$ and $G_i=(G_{i-1})x_i$. Let $|\mathcal{S}|$ be the number of elements in the set $\mathcal{S}$. Let $r_i=|\{y\varepsilon X_{i-1}:(x_i,y)\varepsilon\mathcal{A}_{i-1},G_{i-1}=(X_{i-1},\mathcal{A}_{i-1})\}|$ and $c_i=|\{y\varepsilon X_{i-1}:(y,x_i)\varepsilon\mathcal{A}_{i-1},G_{i-1}=(X_{i-1},\mathcal{A}_{i-1})\}|$ be the out-degree and in-degree, respectively, of vertex x_i in the elimination graph G_{i-1}.

Note that r_i+1, c_i+1 is the number of non-zero elements in the first row, column of the reduced matrix of order n-i+1, i.e. the reduced matrix whose graph is G_{i-1}.

Define $e_{ik}=1$ if there is an arc from x_i to x_k and $e_{ki}=1$ if there is an arc from x_k to x_i in G_{i-1}. Otherwise, let $e_{ik}=0$ and $e_{ki}=0$. We count a division as a multiplication and a subtraction as an addition.

When M is symmetric, then $r_i=c_i\equiv d_i$, and Rose [4] has shown that the factorization $M=LDL^t$ requires $\sum_{i=1}^{n-1} d_i(d_i+3)/2$ multiplications and $\sum_{i=1}^{n-1} d_i(d_i+1)/2$ additions, while the backsolving of $LDL^tx=b$ requires $n+2\sum_{i=1}^{n-1} d_i$ multiplications and $2\sum_{i=1}^{n-1} d_i$ additions.

When M is general, Bunch and Rose [1] have shown that the M=LU factorization requires

$$\sum_{i=1}^{n-1} (r_i+1)c_i \text{ multiplications and}$$

$$\sum_{i=1}^{n-1} r_ic_i \text{ additions, while the}$$

backsolving LUx=b requires

$$n+\sum_{i=1}^{n-1} (r_i+c_i) \text{ multiplications and}$$

$$\sum_{i=1}^{n-1} (r_i+c_i) \text{ additions.}$$

Examples: Let M be an n×n matrix.

(1) If M is symmetric and full, then $d_i=n-i$ and solving $Mx=LDL^tx=b$ requires $n+\sum_{i=1}^{n-1} d_i(d_i+7)/2=\frac{1}{6}n^3+\frac{3}{2}n^2-\frac{2}{3}n$ multiplications and $\sum_{i=1}^{n-1} d_i(d_i+5)/2=\frac{1}{6}n^3+n^2-\frac{7}{6}n$ additions.

(2) If M is symmetric and band, with bandwidth 2m+1, then $d_i=m$ for $1\le i\le n-m$ and $d_i=n-i$ for $n-m+1\le i\le n-1$. Solving $Mx=LDL^tx=b$ requires $(\frac{1}{2}m^2 + \frac{7}{2}m + 1)n - \frac{1}{3}m^3 - 2m^2 - \frac{5}{3}m$ multiplications and $(\frac{1}{2}m^2 + \frac{5}{2}m)n - \frac{1}{3}m^3 - \frac{3}{2}m^2 - \frac{7}{6}m$ additions.

(3) If M is general and full, then $r_i=n-i=c_i$ and solving $Mx=LUx=b$ requires $n+\sum_{i=1}^{n-1}(r_ic_i+2c_i+r_i)=\frac{1}{3}n^3 + n^2 - \frac{1}{3}n$ multiplications and $\sum_{i=1}^{n-1}(r_ic_i+c_i+r_i) = \frac{1}{3}n^3 + \frac{1}{2}n^2 - \frac{5}{6}n$ additions.

(4) If M is general and band, with bandwidth 2m+1, then $r_i=m=c_i$ for $1\le i\le n-m$ and $r_i=n-i=c_i$ for $n-m+1\le i\le n-1$ when no pivoting is needed, while $c_i=m$ for $1\le i\le n-m$, $c_i=n-i$ for $n-m+1\le i\le n-1$, $r_i\le 2m$ for $1\le i\le n-2m$, $r_i\le n-i$ for $n-2m+1\le i\le n-1$, when partial pivoting is used. Solving $Mx=LUx=b$ requires $(m^2+3m+1)n - \frac{2}{3}m^3 - 2m^2 - \frac{4}{3}m$ multiplications and $(m^2+2m)n - \frac{2}{3}m^3 - \frac{3}{2}m^2 - \frac{5}{6}m$ additions when no pivoting is needed, and $\le(2m^2+4m+1)n - \frac{13}{6}m^3 - 4m^2 - \frac{11}{6}m$ multiplications and $\le(2m^2+3m)n - \frac{13}{6}m^3 - \frac{7}{2}m^2 - \frac{4}{3}m$ additions when partial pivoting is used.

3. ERROR ANALYSIS OF TRIANGULAR FACTORIZATION.

Let $||x||_\infty \equiv \max_{1\le i\le n} |x_i|$, $||M||_\infty \equiv \max_{1\le i\le n} \sum_{j=1}^{n} |m_{ij}|$, $||x||_1 \equiv \sum_{i=1}^{n} |x_i|$, $||M||_1 \equiv \max_{1\le j\le n} \sum_{i=1}^{n} |m_{ij}|$. If $M=M^t$, then $||M||_1=||M||_\infty$.

Due to the finite precision arithmetic, we obtain the triangular factorization of a perturbation of M, i.e. $LU=M+F$.

Let $M^{(1)}\equiv M$. Then the elimination is defined sequentially for $1\le k\le n-1$ by the following: given $M^{(k)}$, let $L^{(k)}$ be defined by $\ell_{ii}^{(k)}=1$ for $1\le i\le n$, $\ell_{ik}^{(k)}=f\ell(m_{ik}^{(k)}/m_{kk}^{(k)})$ for $k+1\le i\le n$, $\ell_{ij}^{(k)}=0$ otherwise, then $M^{(k+1)}\equiv f\ell((L^{(k)})^{-1}M^{(k)})$, i.e. $m_{ij}^{(k+1)}\equiv$

$$\begin{cases} 0 & \text{for } k+1\le i\le n,\ j=k \\ f\ell(m_{ij}^{(k)}-\ell_{ik}^{(k)}\times m_{kj}^{(k)}) & \text{for } k+1\le i,\ j\le n \\ m_{ij}^{(k)} & \text{otherwise.} \end{cases}$$

Then $U=M^{(n)}$ and $L=L^{(1)}L^{(2)}\cdots L^{(n-1)}$, i.e.

$$L_{ij} = \begin{cases} 0 & \text{for } i<j \\ 1 & \text{for } i=j \\ L_{ij}^{(j)} & \text{for } i>j \end{cases} .$$

Following the notation of Forsythe and Moler [2], §21, let u be the unit round-off, e.g., $u = \frac{1}{2}\beta^{1-t}$ for rounded operations in t digit, base β arithmetic. Then

$$\ell_{ik}^{(k)} = (m_{ik}^{(k)}/m_{kk}^{(k)})(1+\delta_{ik}e_{ik}), \text{ where } |\delta_{ik}| \leq u, \text{ or}$$

$$0 = m_{ik}^{(k)} - \ell_{ik}^{(k)} m_{kk}^{(k)} + \varepsilon_{ik}^{(k)}, \text{ where } \varepsilon_{ik}^{(k)} \equiv m_{ik}^{(k)} \delta_{ik} e_{ik} \text{ for}$$

$k+1 \leq i \leq n$, and for $k+1 \leq i, j \leq n$:

$$m_{ij}^{(k+1)} = (m_{ij}^{(k)} - \ell_{ik}^{(k)} m_{kj}^{(k)} (1+\delta_{ij}e_{ik}e_{kj}))/(1+\delta'_{ij}e_{ik}e_{kj}),$$

where $\delta'_{11}=0$, $|\delta'_{ii}| \leq u$ for $2 \leq i \leq n$, $|\delta_{ij}| \leq u$ for $1 \leq i, j \leq n$, or

$$m_{ij}^{(k+1)} = m_{ij}^{(k)} - \ell_{ik}^{(k)} m_{kj}^{(k)} + \varepsilon_{ij}^{(k)},$$

where $\varepsilon_{ij}^{(k)} \equiv -\ell_{ik}^{(k)} m_{kj}^{(k)} \delta_{ij} e_{ik} e_{kj} - m_{ij}^{(k+1)} \delta'_{ij} e_{ik} e_{kj}$. Let $\varepsilon_{ij}^{(k)} = 0$ otherwise.

Let $|\ell_{ik}^{(k)}| \leq \tau$ for all i, k and $|m_{ij}^{(k)}| \leq \sigma$ for all i,j,k. Then $|\varepsilon_{ik}^{(k)}| \leq \sigma u e_{ik}$ for $k+1 \leq i \leq n$ and $|\varepsilon_{ij}^{(k)}| \leq (\tau+1)\sigma u e_{ik}e_{kj}$ for $k+1 \leq i, j \leq n$.

Let $F^{(k)} \equiv M^{(k+1)} - M^{(k)} + L^{(k)} M^{(k)}$ for $1 \leq k \leq n-1$;

then $F = \sum_{k=1}^{n-1} F^{(k)}$ and $|F_{ij}| \leq \begin{cases} u(\tau+1)\sigma \sum_{k=1}^{i-1} e_{ik}e_{kj} & \text{for } i \leq j \\ u(\tau+1)\sigma \sum_{k=1}^{j-1} e_{ik}e_{kj} + u\sigma e_{ij} & \text{for } i > j \end{cases}$

Thus, $||F||_\infty \le u\sigma \max_{1\le i\le n}\left\{\sum_{j=1}^{i-1} e_{ij}+(\tau+1)\left[\sum_{j=1}^{i-1}\sum_{k=1}^{j-1} e_{ik}e_{kj}+\sum_{j=i}^{n}\sum_{k=1}^{i-1} e_{ik}e_{kj}\right]\right\}$

and $||F||_1 \le u\sigma \max_{1\le j\le n}\left\{\sum_{i=j+1}^{n} e_{ij}+(\tau+1)\left[\sum_{i=2}^{j}\sum_{k=1}^{i-1} e_{ik}e_{kj}+\sum_{i=j+1}^{n}\sum_{k=1}^{j-1} e_{ik}e_{kj}\right]\right\}$.

If $M=M^t$ and $LDL^t=M+F$, then $F=F^t$, $e_{ij}=e_{ji}$ for all i,j, $|F_{ii}| \le u(\tau+1)\sigma \sum_{k=1}^{i-1} e_{ik}$, $|F_{ij}| \le u(\tau+1)\sigma \sum_{k=1}^{j-1} e_{ik}e_{jk}+u\sigma e_{ij}$ for $i>j$, $|F_{ij}| \le u(\tau+1)\sigma \sum_{k=1}^{i-1} e_{ik}e_{jk}+u\sigma e_{ij}$ for $i<j$, and $||F||_\infty=||F||_1=$

$$\max_{1\le i\le n}\sum_{j=1}^{n}|F_{ij}| \le u\sigma \max_{1\le i\le n}\left\{\sum_{\substack{j=1\\ j\ne i}}^{n} e_{ij}+(\tau+1)\left[\sum_{j=1}^{i-1}\sum_{k=1}^{j-1} e_{ik}e_{jk}+\sum_{k=1}^{i-1} e_{ik}+\sum_{j=i+1}^{n}\sum_{k=1}^{i-1} e_{ik}e_{jk}\right]\right\}.$$

4. ERROR ANALYSIS OF BACKSOLVING.

Solving $Ly=b$, where L is unit lower triangular, we obtain the exact solution of $(L+\delta L)y=b$. Here $y_1=b_1$ and

$$y_i=fl(-\ell_{i1}y_1-\ell_{i2}y_2-\cdots-\ell_{i,i-1}y_{i-1}+b_i)=$$

$$(-\ell_{i1}(1+\delta_{i1})y_1-\cdots-\ell_{i,i-1}(1+\delta_{i,i-1})y_{i-1}+b_i)/(1+\delta_{ii}),$$ where

$\delta_{11}=0$, $|\delta_{ii}|\le u$ for $2\le i\le n$, $|\delta_{i1}|\le 1.01u(1+\sum_{k=2}^{i-1} e_{k1})e_{i1}$ for $2\le i\le n$,

$|\delta_{ij}|\le 1.01u(1+\sum_{k=j+1}^{i-1} e_{kj})e_{ij}$ for $2\le j<i\le n$. Then

$$||\delta L||_1\le 1.01u \max_{i,j}|\ell_{ij}| \max_{1\le j\le n}\{\sum_{i=j}^{n}|\delta_{ij}|\}\le$$

$$1.01u\tau\max\{\tfrac{1}{2}c_1(c_1+1),\ \max_{2\le j\le n-1}[\tfrac{1}{2}c_j(c_j+1)+1]\}$$ since

$$\sum_{i=1}^{n}|\delta_{i1}|\le 1.01u\sum_{i=2}^{n}(1+\sum_{k=2}^{i-1}e_{k1})e_{i1}=1.01uc_1(c_1+1)/2$$ and

$$\sum_{i=j}^{n}|\delta_{ij}|\le 1.01u\{1+\sum_{i=j+1}^{n}(1+\sum_{k=j+1}^{n}e_{kj})e_{ij}\}=1.01u[\tfrac{1}{2}c_j(c_j+1)+1]$$

for $2\le j\le n-1$, while $||\delta L||_\infty\le 1.01u\tau \max_{1\le i\le n}\{e_{i1}+\sum_{k=2}^{i-1}e_{k1}e_{i1}$

$$+\sum_{j=2}^{i-1}(1+\sum_{k=j+1}^{i-1}e_{kj})e_{ij}+1\}\le 1.01u\tau\{1+c_1-1+\sum_{j=2}^{n-1}c_j+1\}=1.01u\tau(1+\sum_{j=1}^{n-1}c_j).$$

Solving $Ux=y$, where U is upper triangular, we obtain the exact solution of $(U+\delta U)x=y$. Here $x_n=fl(y_n/u_{nn})=$

$y_n/[u_{nn}(1+\delta_{nn})]$ and $x_i=fl\left(\dfrac{-u_{i,i+1}x_{i+1}-\cdots-u_{in}x_n+y_i}{u_{ii}}\right)=$

$$\frac{-u_{i,i+1}(1+\delta_{i,i+1})x_{i+1} - \cdots - u_{in}(1+\delta_{in})x_n + y_i}{u_{ii}(1+\delta_{ii})(1+\delta'_{ii})} \quad \text{for } 2 \le i \le n,$$

where $\delta'_{nn}=0$, $|\delta'_{ii}| \le u$ for $1 \le i \le n-1$, $|\delta_{ii}| \le u$ for $1 \le i \le n$,

$|\delta_{in}| \le 1.01u(1+\sum_{k=i+1}^{n-1} e_{kn})e_{in}$ for $1 \le i \le n-1$, and

$|\delta_{ij}| \le 1.01u(2+\sum_{k=i+1}^{j-1} e_{kj})e_{ij}$ for $1 \le i < j \le n-1$. Then

$$||\delta U||_1 \le \max_{i,j}|u_{ij}| \max_{1 \le j \le n} \{|\delta'_{jj}| + \sum_{i=1}^{j} |\delta_{ij}|\} \le$$

$$1.01u\sigma \max\{\max_{1 \le j \le n-1} [2+\sum_{i=1}^{j-1}(2+\sum_{k=i+1}^{j-1} e_{kj})e_{ij}],\ 1+\sum_{i=1}^{n-1}(1+\sum_{k=i+1}^{n-1} e_{kn})e_{in}\},$$

while $||\delta U||_\infty \le \max_{i,j}|u_{ij}| \max_{1 \le i \le n} \{|\delta'_{ii}| + \sum_{j=i}^{n} |\delta_{ij}|\} \le$

$$1.01u\sigma \max_{1 \le i \le n-1} \{2+\sum_{j=i+1}^{n-1}(2+\sum_{k=i+1}^{j-1} e_{kj})e_{ij} + (1+\sum_{k=i+1}^{n-1} e_{kn})e_{in}\} \le$$

$$1.01u\sigma \max_{1 \le i \le n-1} \{2+2r_i-2e_{in}+\tfrac{1}{2}r_i(r_i-1)+e_{in}\} \le$$

$$1.01u\sigma \max_{1 \le i \le n-1} \{2+\tfrac{1}{2}r_i(r_i+3)\}.$$

When $M=M^t$, $U=DL^t$, $r_i=c_i=d_i$ and we obtain $(D+\delta D)z=y$ and $(L^t+\Delta L^t)x=z$. Here $z_i=fl(y_i/d_{ii})=y_i/[d_{ii}(1+\delta_{ii})]$ where $|\delta_{ii}| \le u$ for $1 \le i \le n$, so $||\delta D||_1=||\delta D||_\infty \le u \max_i |d_{ii}| = u\rho$, where $\rho=\max_i |d_{ii}|$.

$$||\Delta L^t||_1 \leq 1.01u\tau \max\{ \max_{1\leq j\leq n-1} [1+\sum_{i=1}^{j-1}(1+\sum_{k=i+1}^{j-1} e_{kj}], \sum_{i=1}^{n-1}\sum_{k=i+1}^{n-1} e_{kn}e_{in}\}$$

and $||\Delta L^t||_\infty \leq 1.01u\tau \max_{1\leq i\leq n-1} \{1+\frac{1}{2}d_i(d_i+1)\}$.

5. ERROR ANALYSIS OF ELIMINATION.

Solving $Mx=b$ in finite precision, we obtain the exact solution of $(M+\delta M)x=b$, where $\delta M=F+(\delta L)U+L(\delta U)+(\delta L)(\delta U)$, or when $M=M^t$, $\delta M=F+LD(\Delta L^t)+[L(\delta D)+(\delta L)D+(\delta L)(\delta D)][L^t+\Delta L^t]$.

$$||L||_1 \le 1+\tau \max_{1\le j\le n-1} c_j, \quad ||L||_\infty \le 1+\tau \max_{2\le i\le n} \sum_{j=1}^{i-1} e_{ij} \le 1+(n-1)\tau,$$

$$||U||_1 \le \sigma(1+\max_{1\le i\le n-1} \sum_{j=i+1}^{n} e_{ij}) \le n\sigma, \quad ||U||_\infty \le \sigma(1+\max_{1\le j\le n-1} r_j),$$

$$||L^t||_1 = ||L||_\infty, \quad ||L^t||_\infty = ||L||_1, \text{ and } ||D||_1 = ||D||_\infty \le \rho.$$

Thus, in general, $||\delta M||_1 \le 1.01u\sigma\Bigg\{\max_{1\le j\le n}\Bigg\{\sum_{i=j+1}^{n} e_{ij}+$

$$(\tau+1)\left[\sum_{i=2}^{j}\sum_{k=1}^{i-1} e_{ik}e_{kj}+\sum_{i=j+1}^{n}\sum_{k=1}^{j-1} e_{ik}c_{kj}\right]\Bigg\}$$

$$+\tau \max_{1\le j\le n-1}\left[1+\tfrac{1}{2}c_j(c_j+1)\right]\left[1+\max_{1\le i\le n-1}\sum_{j=i+1}^{n} e_{ij}\right]$$

$$+\max\Bigg\{\max_{1\le j\le n-1}\left[2+\sum_{i=1}^{j-1}\left(2+\sum_{k=i+1}^{j-1} e_{kj}\right)e_{ij}\right],$$

$$1+\sum_{i=1}^{n-1}\left(1+\sum_{k=i+1}^{n-1} e_{kn}\right)e_{in}\Bigg\}\Bigg(1+\tau \max_{1\le j\le n-1} c_j$$

$$+1.01u\tau \max_{1\le j\le n-1}\left[1+\tfrac{1}{2}c_j(c_j+1)\right]\Bigg)\Bigg\}. \quad ||\delta M||_\infty \text{ can be bounded}$$

similarly. Similar bounds are obtainable when $M=M^t$.

6. THE BAND AND FULL CASES.

When M is a band matrix with bandwidth 2m+1, then $e_{ij}=0$ for $|i-j|>m$; under Gaussian elimination with partial pivoting, $e_{ij}=0$ for $i>j+m$ and for $i<j+2m$.

With no pivoting, $||F||_\infty \le \sigma u[m+(\tau+1)m^2]$, $||L||_\infty \le 1+m\tau$, $||\delta L||_\infty \le 1.01u\tau[\frac{1}{2}m^2+\frac{1}{2}m+1]$, $||U||_\infty \le (m+1)\sigma$, and $||\delta U||_\infty \le 1.01u\sigma[\frac{1}{2}m^2+\frac{3}{2}m+2]$; thus $||\delta M||_\infty \le 1.01u\sigma\{\tau m^3+\frac{1}{2}(5\tau+3)m^2+\frac{1}{2}(7\tau+5)+\tau+2+1.01u\tau(\frac{1}{4}m^4+m^3+\frac{9}{4}m^2+\frac{5}{2}m+2)\}$.

If M is diagonally dominant, then $\tau \le 1$, $\sigma \le 2$, and $||\delta M||_\infty \le 1.01u\{2m^3+8m^2+12m+4+1.01u(\frac{1}{2}m^4+2m^3+\frac{9}{2}m^2+5m+4)\}$.

With partial pivoting, $\tau=1$, $\sigma \le 2^{2m-1}-(m-1)2^{m-2}$ (Wilkinson[5]), $||F||_\infty \le u\sigma[5m^2+4m]$, $||L||_\infty \le 1+m\tau$, $||\delta L||_\infty \le 1.01u[\frac{1}{2}m^2+\frac{1}{2}m+1]$, $||U||_\infty \le (2m+1)\sigma$, and $||\delta U||_\infty \le 1.01u\sigma[2m^2+3m+2]$; thus $||\delta M||_\infty \le 1.01u\sigma\{3m^3+\frac{23}{2}m^2+\frac{23}{2}m+3+1.01u(m^4+\frac{5}{2}m^3+\frac{9}{2}m^2+4m+2)\}$.

When M is full, $e_{ij}=1$ for all i,j. Thus $||F||_\infty \le u\sigma\{n-1+(\tau+1)(\frac{1}{2}n^2-\frac{1}{2}n)\}=\frac{1}{2}u\sigma n[(\tau+1)n-\tau+1]$, $||L||_\infty \le 1+(n-1)\tau$, $||U||_\infty \le n\sigma$, $||\delta L||_\infty \le 1.01u\tau[n(n-1)/2+1]$, $||\delta U||_\infty \le 1.01u\sigma n(n+1)/2$, and $||\delta M||_\infty \le 1.01u\sigma[\tau n^3+n^2+(2-\tau)n+1.01u\tau(n^4+2n^2+2n)/4]$.

With partial or complete pivoting, $\tau=1$ and $||\delta M||_\infty \le 1.01u\sigma[n^3+n^2+n+1.01u(n^4+2n^2+2n)/4]$.

7. REMARKS.

From §§3-5 we see that the error matrix arising in sparse elimination depends on the fill-in occurring during elimination. In the band case we know a priori the structure of the matrix. However, in the arbitrarily sparse case we must view elimination in the context of optimal ordering, i.e. an ordering of the equations so that fill-in is minimized. Given a fixed ordering, our analysis here bounds the errors occurring in the computation.

REFERENCES

[1] Bunch, J.R., and D.J. Rose, "Partitioning, tearing, and modification of sparse linear systems", Cornell University TR 72-149, submitted to Linear Algebra and its Applications.

[2] Forsythe, G.E., and C.B. Moler, Computer Solution of Linear Algebraic Systems, Prentice-Hall, 1967.

[3] George, J.A., "Nested dissection of a regular finite element mesh", SIAM Numerical Analysis, to appear.

[4] Rose, D.J., "A graph-theoretic study of the numerical solution of sparse positive definite systems of linear equations", Graph Theory and Computing, R. Read, editor, Academic Press, New York, 1972.

[5] Wilkinson, J.H., private communication.